Direct detection of DNA methylation during single-molecule, real-time sequencing

Benjamin A Flusberg, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson A Clark, Jonas Korlach & Stephen W Turner

We describe the direct detection of DNA methylation, without bisulfite conversion, through single-molecule, real-time (SMRT) sequencing. In SMRT sequencing, DNA polymerases catalyze the incorporation of fluorescently labeled nucleotides into complementary nucleic acid strands. The arrival times and durations of the resulting fluorescence pulses yield information about polymerase kinetics and allow direct detection of modified nucleotides in the DNA template, including N6-methyladenine, 5-methylcytosine and 5-hydroxymethylcytosine. Measurement of polymerase kinetics is an intrinsic part of SMRT sequencing and does not adversely affect determination of primary DNA sequence. The various modifications affect polymerase kinetics differently, allowing discrimination between them. We used these kinetic signatures to identify adenine methylation in genomic samples and found that, in combination with circular consensus sequencing, they can enable single-molecule identification of epigenetic modifications with base-pair resolution. This method is amenable to long read lengths and will likely enable mapping of methylation patterns in even highly repetitive genomic regions.

DNA methylation, in its various forms, has been implicated in the regulation of a variety of biological processes across virtually every branch of the taxonomic tree. For example, in certain bacteria, N6-methyladenine (mA) appears primarily in GATC sequence contexts and helps regulate replication, the mismatch repair pathway and the expression of certain genes[1]. In plants, 5-methylcytosine (mC) appears in multiple sequence contexts, each controlled by separate genetic mechanisms[2,3]. In vertebrates, mC usually occurs in C-G dinucleotides, which often cluster in regions called CpG islands that are at or near transcription start sites[4–6]. Methylation within CpG islands regulates gene expression in cells[7] and can also confer epigenetic heritability in offspring[8,9]. Changes in mC patterns have a crucial role in development[10,11] and have been associated with cancer[12,13] and other diseases[14]. Abundant cytosine methylation in non–C-G contexts has been recently found in human embryonic stem cells but not in differentiated cells, suggesting that it is a distinct type of methylation involved in the maintenance of the pluripotent state[15]. Finally, 5-hydroxymethylcytosine (hmC) is a newly identified epigenetic mark whose biological function is not yet understood, found thus far in mouse Purkinje neurons[16] and embryonic stem cells[17].

Because of its key role in human health and disease, cytosine methylation is the most widely studied of the DNA modifications described above, and there is much interest in mapping genome-wide mC patterns across different cell types and in response to various environmental influences[18]. Currently, the most common technique for studying cytosine methylation involves bisulfite treatment (which transforms epigenetic information to genetic information by converting cytosine, but not methylcytosine, to uracil) followed by massively parallel DNA sequencing[18]. Using this approach, researchers have recently constructed single-base resolution methylation maps for the *Arabidopsis thaliana* genome[2,3], for a subset of the mouse genome[19] and for both fibroblasts and embryonic stem cells throughout the majority of the human genome[15]. Despite these advances, enabled by bisulfite treatment–based sequencing, there are several drawbacks to the technique. For example, the sample preparation steps associated with bisulfite treatment can be costly and time-consuming, and the harsh reaction conditions necessary for complete conversion can degrade DNA. Additionally, the reduction of complexity in converted genomes constrains primer design for subsequent PCR amplification[20] and complicates alignment to a reference genome[6]. Finally, discrimination between cytosine, mC and hmC cannot be accomplished with bisulfite treatment–based sequencing[16,17,21,22].

Direct detection of methylation is possible for nucleotides in solution using techniques such as thin layer chromatography[16,17], high-performance liquid chromatography[16,23], mass spectrometry[16,17,23] and nanopore amperometry[24]. To our knowledge, however, no high-throughput method has been demonstrated that allows the determination of primary sequence at the same time as methylation status. Here we present a method to directly detect DNA methylation during single-molecule, real-time (SMRT) DNA sequencing, a technique for studying nucleic acid sequence and structure[25]. In this technique, single DNA polymerase molecules are observed in real time while they catalyze the incorporation of fluorescently labeled nucleotides complementary to a template nucleic acid strand. These reactions are measured

Pacific Biosciences, Menlo Park, California, USA. Correspondence should be addressed to S.W.T. (sturner@pacificbiosciences.com).

RECEIVED 31 DECEMBER 2009; ACCEPTED 8 APRIL 2010; PUBLISHED ONLINE 9 MAY 2010; DOI:10.1038/NMETH.1459

simultaneously in thousands of arrayed zero-mode waveguides (ZMWs)[26], nanophotonic structures that reduce background fluorescence, thereby enabling use of the high concentrations of labeled nucleotides necessary to support fast and processive DNA sequencing by synthesis. Incorporation of a nucleotide is detected as a pulse of fluorescence whose color identifies that nucleotide. The pulse ends when the fluorophore, linked to the nucleotide's terminal phosphate, is cleaved by the polymerase before translocation to the next base in the DNA template. Typical synthesis rates in SMRT sequencing are 1–3 bases per second in the system used here[25].

Fluorescence pulses in SMRT sequencing are characterized not only by their emission spectra but also by their duration and by the interval between successive pulses[25]. These metrics, defined here as pulse width and interpulse duration (IPD), add valuable information about DNA polymerase kinetics. Pulse width is a function of all kinetic steps after nucleotide binding and up to fluorophore release, and IPD is determined by the kinetics of nucleotide binding and polymerase translocation[27]. We have previously demonstrated that SMRT sequencing polymerase synthesis rates are sensitive to DNA primary and secondary structure[25]. Therefore, we hypothesized that methylated bases in a DNA template might be detected directly on the principle that their presence affects polymerase kinetics during SMRT sequencing (**Fig. 1**). Curiously, to our knowledge, the kinetics of nucleotide incorporation opposite methylated templates have not been studied previously, even in bulk, despite evidence that other types of modified nucleotides alter DNA polymerase kinetics[28].

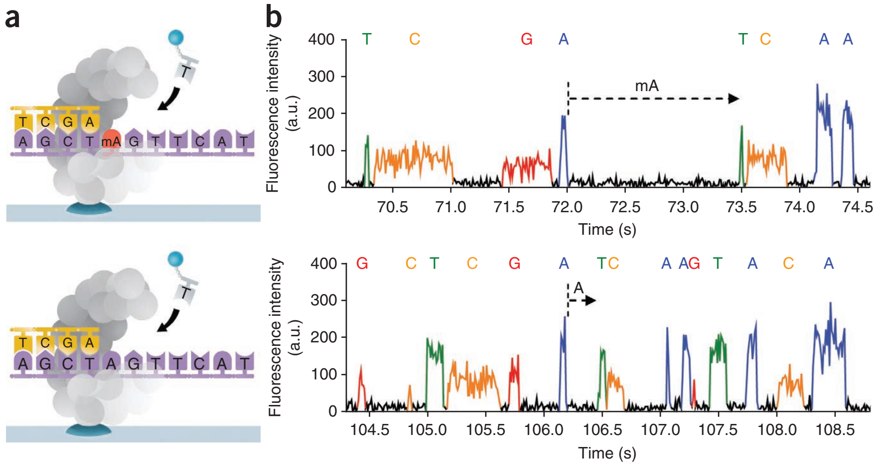

Figure 1 | Principle and corresponding example of detecting DNA methylation during SMRT sequencing. (**a**) Schematics of polymerase synthesis of DNA strands containing a methylated (top) or unmethylated (bottom) adenine. (**b**) Typical SMRT sequencing fluorescence traces for samples in **a**. Letters above the fluorescence trace pulses indicate the identity of the nucleotide incorporated into the growing complementary strand. Dashed arrows indicate the IPD before incorporation of the cognate thymine. For this typical example, the IPD is about five times larger for mA in the template compared to adenine.

RESULTS

Effects of methylation on polymerase kinetics

To test this hypothesis, we designed several synthetic DNA templates that were identical except for their methylation status at specific sites. The control template contained no methylation, and other templates contained several mA, mC or hmC bases. Because the methylated bases could, in principle, affect the kinetics of DNA synthesis over a range of several nearby bases, we separated them by no less than 11 bases[29]. In all cases, mA was located in a GATC sequence context, and mC and hmC were located in C-G contexts. We sequenced these templates and then compared the average IPD in each of the methylated templates to the average IPD in the control template by computing their ratio at every template position. For all three methylation types, there was a clear excursion in polymerase synthesis kinetics in the vicinity of the methylated bases (**Fig. 2**). The methylated base is in contact with the polymerase for several bases before and after occupying the polymerase active site[29]. Consistent with this model, the kinetic impact of methylation is not restricted to the nucleotide incorporation opposite the modified base. Additionally, the IPD ratio patterns differed for the two methylated positions. As the only source of difference between these two loci in all three template types

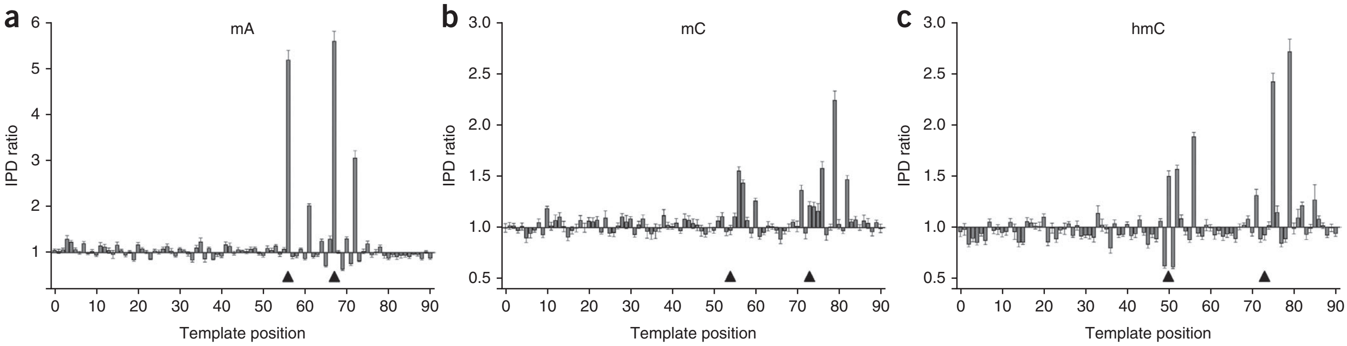

Figure 2 | SMRT sequencing–mediated detection of methylated DNA bases. (**a**–**c**) Ratios of the average IPD in the methylated template to the average IPD in the control template, plotted versus DNA template position. In the region shown, the two templates are identical except at the two positions marked by triangles, which are mAs in **a**, mCs in **b** and hmCs in **c** for the methylated templates. Polymerase synthesis proceeds in the direction of increasing position number. The templates have a circular topology and are 199 bases long, but only 90-base segments surrounding the methylated regions are shown for clarity. Error bars, s.e.m.; *n* = 346 measurements for each template position in **a**, *n* = 504 in **b** and *n* = 393 in **c**, computed using bootstrapping techniques.

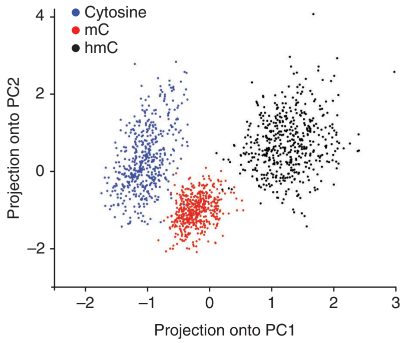

Figure 3 | Principal component analysis of cytosine, mC and hmC IPD and pulse width signatures. Each principal component is a linear combination of the mean IPD and pulse width at positions 71–79, which surround the variably modified template position 73. The weightings of IPD and pulse width at each position are shown in **Supplementary Table 1**. Data points on the plot were computed by projecting a random 20% subsample of the IPD and pulse width values onto the first two principal components (PC1 and PC2) and then converting to a *z* score.

was the local sequence context (template sequences are listed in **Supplementary Note 1**), we conclude that, in general, the kinetic signatures of methylation will be sequence context–dependent.

There were also several common features for the instances of each methylation type in each template, suggesting that there may be universal interactions between the methylated nucleobase and specific sites of the polymerase. For mA (**Fig. 2a**), the ratio of IPDs was greatest (ranging between 5 and 6) opposite the methylated positions themselves. The N6 position is involved in hydrogen bonding during complementary base pairing, and therefore it is possible that the methyl group of mA directly modifies nucleotide binding kinetics. Another characteristic common to the two mA positions is an excursion in the IPD ratio five bases after incorporation opposite the methylated base.

The templates containing mC (**Fig. 2b**) had considerable IPD increases two, three and six bases after both methylated positions. The templates containing hmC (**Fig. 2c**) exhibited similar IPD signals at positions two and six, but not three, bases after the methylated bases. Plots of pulse width ratios had more pronounced excursions for hmC than for mC (**Supplementary Fig. 1**). In fact, because each modification had a unique IPD and pulse width signature in a given context (position 73, for example; **Fig. 2b**,**c** and **Supplementary Fig. 1**), this approach opens up the intriguing possibility of directly distinguishing between cytosine, mC and hmC during real-time sequencing. To this end, we used principal component analysis to find the combination of weights for the IPD and pulse width signals at the various template positions near the putative modification that optimizes the resolution (**Supplementary Table 1**). The separation between projections of the kinetic signatures for each template onto the first two principal components (**Fig. 3**) demonstrated the discrimination among these three cytosine nucleobase types by using information from multiple kinetic parameters at multiple template positions.

Methylation detection by circular consensus sequencing

Whereas in previous experiments we established the principle of methylation detection in populations of identical molecules, in practice, individual bases at a given position in a genomic sample might be methylated in only a fraction of the molecules. When a base at a given genome position is not always methylated, the aggregate kinetic data over an ensemble will be a linear superposition of the kinetic signatures for the methylated and unmethylated cases. Partial methylation can be quantified by fitting the data to a two-component model, but for the highly overlapping IPD distributions that result from a single, rate-limiting step[25], it can be preferable to sequence a smaller number of molecules but multiple times each. To enable reading of individual molecules multiple times, we exploited the circular topology of our DNA templates, achieved by the ligation of hairpin adaptors to both ends of a double-stranded DNA insert (**Fig. 4a**). A strand-displacing DNA polymerase can carry out multiple laps of DNA synthesis around such a DNA template and enable repeated, or circular consensus, sequencing of the same DNA molecule. Repeated measurements (which we call circular subreads) of IPD at a particular DNA template position yielded a mean IPD at that position that followed a gamma distribution, which is narrower than the underlying exponential distribution. As we collected more circular subreads, the distributions of mean IPD for methylated and unmethylated bases became better separated (**Fig. 4b**). This substantially improved the discrimination between them, as shown by receiver operating characteristics (ROC) curves for calling adenosine versus mA (**Fig. 4c**) in a particular context. The normalized area under the

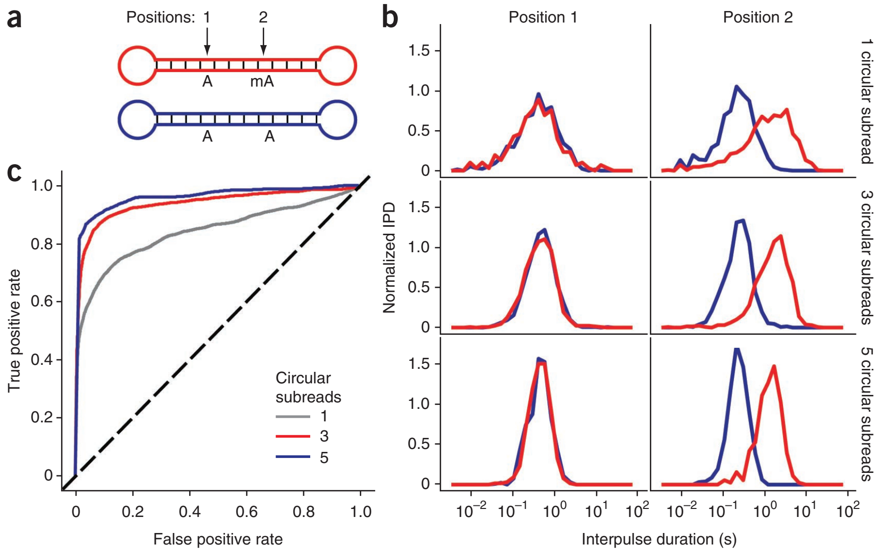

Figure 4 | IPD distributions for adenine and mA in synthetic DNA templates. (**a**) Schematic of the 199-base DNA templates used in this experiment. (**b**) IPD distributions at the indicated positions for both templates, colored as in **a**. The histograms depict the distributions of mean IPD (averaged over the indicated number of circular subreads). (**c**) Receiver operating characteristic curves, based on the IPD distributions from the differentially methylated position in **b** and parameterized by IPD threshold, for assigning a methylation status to an adenine nucleotide after 1, 3 and 5 circular consensus sequencing subreads. The dashed line depicts the receiver operating characteristic curve for randomly guessing the methylation status.

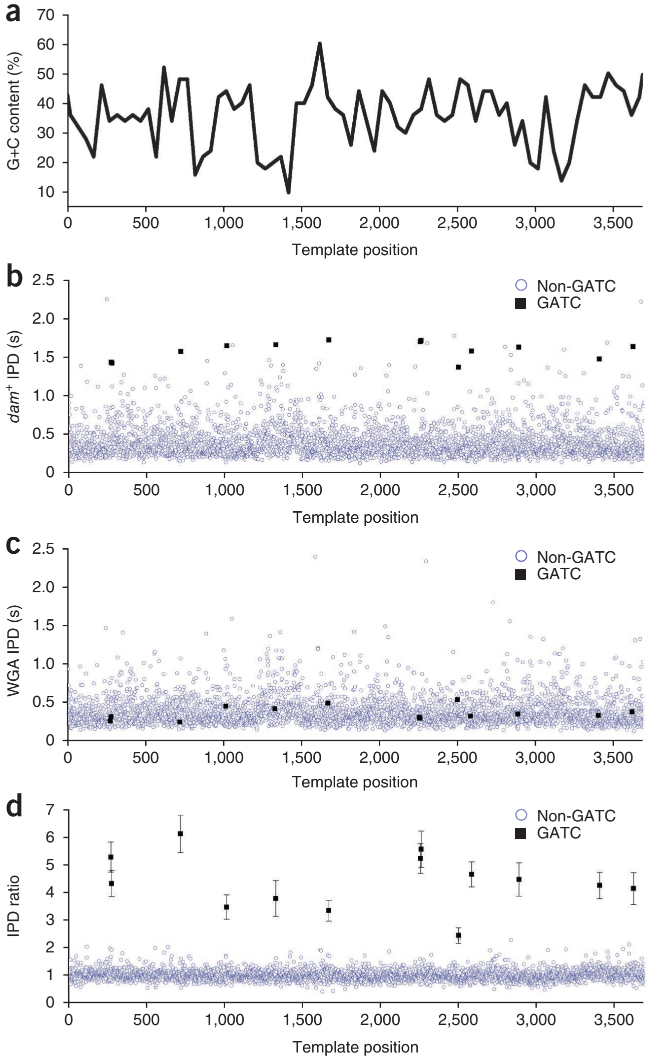

Figure 5 | Comparison of SMRT sequencing kinetics for DNA samples propagated in *dam*+ *E. coli* and for the same samples after whole-genome amplification. (**a**) A 50-bp-window G+C-content of the sample, plotted versus template position. (**b**,**c**) Average IPD at each template position in the *dam*+ sample (**b**) and in the WGA sample (**c**). (**d**) Ratio of the average IPDs (for *dam*+ sample in **b** divided by that for WGA sample in **c**). Positions with a GATC context, where methylation of adenine at the sequence motif GATC is expected, are denoted by black squares, and all other positions are denoted by open blue circles. Error bars at the GATC positions denote the s.e.m. IPD ratio at those positions (n = 106 measurements at each position). For comparison, the mean ± s.d. of all IPD ratios at non-GATC positions was 1.00 ± 0.24 (n = ~389,000 measurements). Average sequencing coverage across this fosmid region was 121-fold for the *dam*+ sample and 91-fold for the WGA sample.

ROC curve was 0.80 after the first circular subread but increased to 0.92 and 0.96 after three and five circular subreads, respectively. After five subreads, >85% of mA bases could be detected at this template position with a false positive rate of only ~5%. Because the templates each have a length of 199 bases, five full circular subreads correspond to read lengths of nearly 1,000 bases. Additional subreads enabled by longer read lengths would yield even better discrimination. The discrimination can be better still if all of the template positions affected by the methylated site are taken into consideration (as we observed with mC and hmC; **Fig. 3**). Notably, there was a similarity between using multiple observations of the same template position through circular consensus and using multiple affected positions from the same subread. Both led to transitions from highly overlapping exponential distributions to better-separated modal distributions. The combination of circular consensus sequencing and methods (such as principal component analysis) to combine all available information will greatly aid in the extension of this technique to quantification of variably methylated genomic sites with base-pair resolution.

Adenine methylation in *Escherichia coli*

To test direct methylation detection with genomic DNA, we mapped the dependence of IPD on sequence context for a *Caenorhabditis elegans* fosmid, isolated from a DNA adenine methyltransferase–positive (*dam*+) *E. coli* strain. To obtain an unmethylated control template, we subjected a portion of the sample to whole-genome amplification (WGA), which we expected to erase any methylation signatures (**Supplementary Fig. 2**). We examined the sequencing kinetics of a 3.7-kb subregion of this fosmid (**Fig. 5**, **Supplementary Fig. 3** and **Supplementary Data**). In a range of sequence contexts with varying G+C content (**Fig. 5a**), the *dam*+ samples had average IPDs at GATC positions that were generally greater than those at non-GATC positions (**Fig. 5b**). In contrast, average IPDs were similar at all template positions in the WGA samples (**Fig. 5c**). The ratio of average IPDs between the two samples (**Fig. 5d**) demonstrates that polymerase kinetics are altered substantially by adenine methylation in various sequence contexts (**Table 1**). The IPD ratio increase was similar over different G+C-content levels in this sample (**Supplementary Fig. 4**). The increase in average IPD caused by adenine methylation in *E. coli* was consistent with the range of IPD ratios measured with the synthetic mA templates. Average IPDs over all possible 4-mer sequence contexts in the entire fosmid sample (48 kb including the vector; **Supplementary Fig. 5**) had a notable context dependence, evident as a nonrandom profile in the *dam*+ and WGA sample heatmaps, highlighting the sensitivity of the method for studying DNA polymerase kinetics. The high degree of similarity between the two maps demonstrates the robustness of SMRT sequencing

Table 1 | IPD ratios for each fosmid GATC motif

Position	Sequence	IPD ratio	*P* value
273	TGCCATGATCTAGATC	5.28	2.84×10^{-18}
279	GATCTAGATCATCGTG	4.32	4.41×10^{-16}
720	TTCTATGATCAGGGAG	6.12	7.00×10^{-21}
1015	GCGTGGGATCTGTATG	3.46	6.25×10^{-13}
1329	TATCACGATCTCATTA	3.78	2.35×10^{-12}
1668	TAGTTGGATCAAGAGA	3.33	2.61×10^{-13}
2256	CTTTTGGATCAGATCC	5.22	1.70×10^{-19}
2261	GGATCAGATCCAATTA	5.56	1.43×10^{-22}
2499	CAGATGGATCAATCAA	2.43	4.44×10^{-7}
2583	ATTTTTGATCTAGTTT	4.65	2.00×10^{-21}
2887	ATTCGCGATCTCCACA	4.46	1.57×10^{-14}
3402	CCTCAAGATCATCATC	4.24	2.04×10^{-15}
3619	GCCAGCGATCATATTT	4.13	2.05×10^{-10}

For each GATC motif in the 3.7-kb fosmid subregion (**Fig. 5**), the local sequence context, IPD ratio (average IPD for the *dam*+ sample divided by the average IPD for the WGA sample) and *P* value are shown. The *P* value was derived by performing a two-sample Kolmogorov-Smirnov goodness-of-fit test, which compares the IPD data at each position in the *dam*+ and WGA samples and tests the likelihood that they are drawn from the same underlying distribution (the null hypothesis). Lower *P* values indicate greater confidence that the null hypothesis should be rejected.

IPD measurements. The notable exception to their similarity is for the GATC sequence context, which has a mean IPD about four times greater in the *dam*+ samples than in the WGA samples. Extension to much larger genomic samples will be straightforward using future commercial versions of SMRT sequencing instrumentation that have ~100 times greater throughput than the prototype instrument[25,30] used in these experiments.

DISCUSSION

In SMRT DNA sequencing, polymerase kinetics are measured alongside primary sequence determination and require no additional sample preparation steps. We showed that mA, mC and hmC in a DNA template alter incorporation kinetics and expect that other epigenetic modifications, as well as various forms of DNA damage, may also be detected by this method. Unique kinetic signatures displayed by each modification will permit discrimination between them in the same DNA sample. By enabling repeated interrogation of individual molecules, circular consensus sequencing allows base-pair resolution and single-molecule sensitivity for detection of mA. For mC and hmC, enhancements of kinetic sensitivity will likely be required. Such improvements could come from optimized solution conditions, polymerase mutations and algorithmic approaches that take advantage of the kinetic signatures' spread over multiple template positions, and deconvolution techniques will help resolve neighboring mC bases (**Supplementary Fig. 6**). We sequenced both the methylated and control DNA for each experiment described here, but in resequencing applications, unmethylated kinetic reference data could be collected just once and tabulated for all subsequent studies of the same species. In the future, *de novo* detection of methylation may also be possible by tabulating expected kinetics over a suitable number of contexts or by taking advantage of heuristics that embody the observed trends in SMRT sequencing kinetics.

The long read lengths of SMRT sequencing will likely permit methylation profiling in highly repetitive genomic regions, in which a substantial fraction of mC residues resides[6]. Combined with single-molecule sensitivity, these long reads will also allow phasing of methylation status between different genomic positions. As we continue to refine this technique, *de novo* methylation profiling may become possible.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank the entire staff at Pacific Biosciences, in particular J. Londry and D. Kolesnikov for sample preparation; E. Mollova, M. Berhe and J. Yen for running sequencing experiments; J. Sorenson, J. Chin, A. Kislyuk and D. Holden for help with data analysis; and E. Schadt and J. Eid for helpful discussions. This work was supported by US National Human Genome Research Institute grant 1RC2HG005618-01.

AUTHOR CONTRIBUTIONS

B.A.F., K.J.T., J.K., J.H.L. and S.W.T. designed the experiments. E.C.O. and T.A.C. prepared fosmid library constructs. B.A.F. conducted the sequencing experiments. D.R.W. and B.A.F. analyzed data. B.A.F., J.K., S.W.T., E.C.O., D.R.W. and T.A.C. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Marinus, M.G. & Casadesus, J. Roles of DNA adenine methylation in host-pathogen interactions: mismatch repair, transcriptional regulation, and more. *FEMS Microbiol. Rev.* **33**, 488–503 (2009).
2. Cokus, S.J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
3. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
4. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
5. Saxonov, S., Berg, P. & Brutlag, D.L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. USA* **103**, 1412–1417 (2006).
6. Pomraning, K.R., Smith, K.M. & Freitag, M. Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods* **47**, 142–150 (2009).
7. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* **33** (Suppl.), 245–254 (2003).
8. Holliday, R. & Pugh, J.E. DNA modification mechanisms and gene activity during development. *Science* **187**, 226–232 (1975).
9. Riggs, A.D. X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell Genet.* **14**, 9–25 (1975).
10. Li, E., Bestor, T.H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915–926 (1992).
11. Razin, A. & Shemer, R. DNA methylation in early development. *Hum. Mol. Genet.* **4** Spec No, 1751–1755 (1995).
12. Jones, P.A. & Baylin, S.B. The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* **3**, 415–428 (2002).
13. Jones, P.A. & Laird, P.W. Cancer epigenetics comes of age. *Nat. Genet.* **21**, 163–167 (1999).
14. Robertson, K.D. DNA methylation and human disease. *Nat. Rev. Genet.* **6**, 597–610 (2005).
15. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
16. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in purkinje neurons and the brain. *Science* **324**, 929–930 (2009).
17. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
18. Lister, R. & Ecker, J.R. Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Res.* **19**, 959–966 (2009).
19. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).
20. Clark, S.J., Statham, A., Stirzaker, C., Molloy, P.L. & Frommer, M. DNA methylation: bisulphite modification and analysis. *Nat. Protocols* **1**, 2353–2364 (2006).
21. Hayatsu, H. & Shiragami, M. Reaction of bisulfite with the 5-hydroxymethyl group in pyrimidines and in phage DNAs. *Biochemistry* **18**, 632–637 (1979).
22. Huang, Y. *et al.* The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS One* **5**, e8888 (2010).
23. Tardy-Planechaud, S., Fujimoto, J., Lin, S.S. & Sowers, L.C. Solid phase synthesis and restriction endonuclease cleavage of oligodeoxynucleotides containing 5-(hydroxymethyl)-cytosine. *Nucleic Acids Res.* **25**, 553–559 (1997).
24. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009).
25. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
26. Levene, M.J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682–686 (2003).
27. Wong, I., Patel, S.S. & Johnson, K.A. An induced-fit kinetic mechanism for DNA replication fidelity: direct measurement by single-turnover kinetics. *Biochemistry* **30**, 526–537 (1991).
28. Hsu, G.W., Ober, M., Carell, T. & Beese, L.S. Error-prone replication of oxidatively damaged DNA by a high-fidelity DNA polymerase. *Nature* **431**, 217–221 (2004).
29. Berman, A.J. *et al.* Structures of phi29 DNA polymerase complexed with substrate: the mechanism of translocation in B-family polymerases. *EMBO J.* **26**, 3494–3505 (2007).
30. Lundquist, P.M. *et al.* Parallel confocal detection of single molecules in real time. *Opt. Lett.* **33**, 1026–1028 (2008).

ONLINE METHODS

Zero-mode waveguide (ZMW) fabrication. ZMW nanostructures were fabricated and functionalized as previously described[25,31,32]. Sequencing experiments were performed using arrays of 3,000 ZMWs monitored simultaneously[25,31,32].

Preparation of DNA templates. Sets of ~35-base single-stranded DNA oligonucleotides (for experiments described in **Figs. 1–4** and **Supplementary Fig. 1**) were purchased (Trilink Biotechnologies). Presence of base modifications in these single-stranded oligonucleotides was verified by mass spectrometry. After hybridization and ligation, each end of the resulting double-stranded DNA oligonucleotides was ligated to a hairpin oligonucleotide. Samples were treated with exonucleases to remove any molecules that were not covalently closed. Sequences for the resulting DNA templates, which were 199 bases long and consisted of a central 84-bp double-stranded region with single-stranded loops at each end, are shown in **Supplementary Note 1**.

For sequencing of the full fosmid (**Supplementary Figs. 2** and **5**), a fosmid clone (clone identifier WRM0639cE06) containing an ~40 kb *C. elegans* genomic insert was obtained from Geneservice (http://www.geneservice.co.uk/products/clones/Celegans_Fos.jsp) in *dam*$^+$ *E. coli* strain EPI300, and grown and amplified using the inducible origin (CopyControl system; Epicentre). Fosmid DNA was purified using standard methods. DNA templates were then created directly from fosmid DNA or from whole-genome amplified fosmid DNA (WGA sample). For WGA libraries, 25 ng of fosmid DNA was amplified using the manufacturer-recommended conditions in the GenomiPhi HY DNA Amplification kit (GE Healthcare).

For sequencing the subsection of the fosmid (**Fig. 5** and **Supplementary Figs. 3** and **4**), an ~3.7 kb segment (corresponding to positions 12797–16484 in the fosmid) containing 13 instances of the GATC sequence context was PCR-amplified from the fosmid using Phusion High-Fidelity DNA polymerase (New England Biolabs) and the following primers: forward, 5′-AGTCCTGATGCTTTCACCAAAT-3′ and reverse, 5′-ATTTAGATTGCCAAAGCCGTAA-3′. PCR products were cloned into the pCR-Blunt vector using the Zero Blunt PCR Cloning kit (Invitrogen) and propagated in the *dam*$^+$ *E. coli* strain TOP10 (Invitrogen). Approximately 25 ng of the DNA was amplified using the REPLI-g Mini Kit (Qiagen) to generate the unmethylated control sample.

Fosmid DNA, or an equivalent quantity of WGA fosmid DNA, was sheared to a mean size of 500 bp (**Fig. 5** and **Supplementary Figs. 3** and **4**) or 200 bp (**Supplementary Fig. 5**) using an ultrasonicator (Covaris). Sheared DNA was then end-repaired with a cocktail of T4 DNA polymerase and T4 polynucleotide kinase, purified and 3′ A-tailed with Klenow(exo–). The A-tailed fragments were ligated to hairpin oligonucleotides that contained a single 3′ thymine overhang and 5′ phosphate. Samples were treated with a mixture of exonucleases to remove any molecules that were not covalently closed. The resulting DNA templates were purified using solid-phase reversible immobilization (SPRI) magnetic beads (AMPure; Agencourt Bioscience) and annealed to a twofold molar excess of a sequencing primer (5′-GGAGGAGGAGGA-3′) that specifically bound to the single-stranded loop region of the hairpin adapters.

Preparation of DNA polymerase and phospholinked dNTP, and DNA sequencing assays. DNA polymerases were generated as described previously[25,33]. Phospholinked dNTPs were generated as described previously[25,33], with the exception of replacing Alexa Fluor 660 with a modified Cy5.5 fluorophore. Additional modifications included permuting the nucleobases associated with each dye to the following configuration: Alexa Fluor 555-dT, Alexa Fluor 568-dG, Alexa Fluor 647-dA and Cy5.5-dC. The excitation laser lines used were the same as described previously[25,30]. Protocols for DNA polymerase/template complex formation, complex immobilization on the ZMW array and sequencing reactions were similar to those described previously[25].

Data collection and analysis. Data were collected on a highly parallel confocal fluorescence detection instrument, as previously described[25,30]. Pulse calling, which used a threshold algorithm on the dye-weighted intensities of fluorescence emissions, and read alignments, achieved using a Smith-Waterman algorithm, have been described prevously[25]. Reads were filtered after alignment to remove low-quality sequences derived from doubly loaded ZMWs. IPD values were tabulated from consecutive pairs of correctly aligning template positions and were assigned to the second template position in the pair. Pulse width values were computed as the duration of the pulses associated with correctly aligning base calls. To avoid outlier effects, the smallest and largest 5% of IPDs and pulse width values at each position were excluded from all analyses.

Bar plot error bars (**Fig. 2**) represent an estimate of the s.e.m. IPD ratio, computed by bootstrapping 10 randomly selected subsamples of 10% of the data. It can be seen that the error bars underestimate the error somewhat, as small excursions do occur at positions far away from any modification, where differences from the control are not expected. Molecular coverage varied between template, with an average coverage of 346-, 504- and 393-fold for the 10% subsamples of the mA, mC and hmC templates, respectively. No bootstrapping was performed for creation of the pulse width plots (**Supplementary Fig. 1**), for which all molecules were used.

Standard principal component analysis[34] was carried out using the prcomp function from the Stats Package of the statistical computing program, R (http://www.R-project.org/). Input variables were scaled to have zero mean and unit variance, and the resulting first and second principal components were determined from the entire dataset (**Supplementary Table 1**). To generate the principal component scatter plot (**Fig. 3**), 500 subsets of 20% of the data for each template were first projected onto these first two principal components. These values were then converted into a *z* score by subtracting the mean and dividing by the s.d. of all 1,500 data points for each principal component.

IPD distributions (**Fig. 4**) were determined by averaging multiple IPD measurements at the same template position within single molecules. All molecules from the mA experiment were used. The corresponding ROC curves were generated by sliding a threshold value across the full range of observed average IPDs. The true positive rate was computed for each threshold as the fraction of methylated observations with an average IPD larger than the threshold. Similarly, the false positive rate was determined by the fraction of nonmethylated observations with an average IPD larger than the threshold.

For the fosmid experiments (**Fig. 5** and **Supplementary Figs. 3–5**), GATC positions are defined as those positions at which a thymine is

doi:10.1038/nmeth.1459

incorporated opposite a template adenine that is in a template GATC context. Non-GATC positions correspond to all other positions. IPD ratios at each position were normalized by the ratio of the average IPD over all *dam*$^+$ reads to the ratio of the average IPD over all WGA sample reads. Average sequencing coverage for the 3,688-bp fosmid region analyzed in **Figure 5** was 121-fold for the *dam*$^+$ sample and 91-fold for the WGA sample. This coverage was obtained using nine ZMW arrays (SMRT cells) and a total of ~1.5 h of sequencing for each sample (*dam*$^+$ and WGA samples).

Local sequence context (**Supplementary Fig. 5**) was determined using the standard Smith-Waterman alignment algorithm. The 'local context' of detected bases was defined as the two bases previously detected, the detected base itself and the next base detected afterward. For example, in the local context 5′-GATC-3′, the mean IPD reported describes the average duration between the detected adenine (complementary to a thymine in the template DNA) and the detected thymine (complementary to an adenine or mA in the template). On average, 1,890 observations (remaining after removal of the smallest and largest 5%) were used to compute the mean IPD for each of the 256 possible 4-mer contexts, corresponding to tenfold coverage of the entire fosmid.

31. Foquet, M. *et al.* Improved fabrication of zero-mode waveguides for single-molecule detection. *J. Appl. Phys.* **103**, 034301 (2008).
32. Korlach, J. *et al.* Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc. Natl. Acad. Sci. USA* **105**, 1176–1181 (2008).
33. Korlach, J. *et al.* Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides. *Nucleosides Nucleotides Nucleic Acids* **27**, 1072–1082 (2008).
34. Jolliffe, I.T. *Principal Component Analysis* 2nd edn. (Springer-Verlag, New York, 2002).

doi:10.1038/nmeth.1459